Pre-Analysis Plan Project Submission 2

Caroline Sullivan, Joe Thompson, Genevieve Purcell, Sasha Porter

Our study's research question is: *How accurately can previous election results predict the outcomes of future elections?* In this study, each observation will be the number of votes a candidate received by party in each presidential election, held every four years, for each county in the state of Virginia. Specifically, we will be looking at whether they win or lose. Since we are using historical data to predict labeled outcomes, we will be conducting a supervised learning study. Our goal is to predict a binary outcome—win or loss—for each candidate or party, making this a classification analysis. Specifically, we aim to determine whether the winning party in previous elections can help forecast performance in the 2024 election.

To analyze this we will use logistic regression to classify outcomes by win or loss for each party. Specifically we will use binary logistic regression because there are a finite number of outcomes: 1 representing a win and 0 representing a loss. By training the model on data points from past elections up until the most recent election, we'll test how well it predicts the upcoming 2024 election outcome. We will know if our approach "works" if we can effectively predict the results of the next election based on previous election data. Success is defined by how accurately we predict the 2024 election results for each party in Virginia. We will then compare it to the actual outcomes of the 2024 presidential election for each party once the results come out.

Challenges could arise from changes in voting behavior, such as increased mail-in voting since 2020, and external influences, such as economic or political shifts. These challenges may not be captured in the data, so we may not be able to directly address these weaknesses. We will still remain aware of their potential impact on our results. If our approach fails, we will learn how to adjust our analysis and fix any easily identifiable problem. In terms of feature engineering, party and mode are categorical variables, so they could be one-hot encoded. To give the logistic regression model more variables to learn patterns from, we can engineer a variable that represents the vote share for each party in each county. This would be done by calculating the percentage of votes each party received in each county per presidential election year.

To prepare our data, we could also sum or average votes across other features besides party, such as demographics (gender, race, socioeconomic background). We will explore creating variables, like historical trends on how party votes change for political parties throughout different election years. Also, we can grab data on demographics for counties and see if counties with more minority voters tend to lean a certain way party wise.

We will communicate our results using a combination of tables and visualizations. This may include maps of Virginia, showing the winning party by county in current and historical contexts. Graphs and visualizations highlighting the historical trend of the winning party by county and any correlations between county demographics and party leaning. A table of regression coefficients will provide insight into how past presidential election data contributes to predicting the current year while controlling for other factors.